IT UNIVERSITY OF COPENHAGEN

Projects in Data Science

Group Seahorse

Emil Sander Korczak - ekor@itu.dk

Joakim Smidsgaard Andersen - smja@itu.dk

Radost Pencheva Boyadzhieva - radb@itu.dk

Sigrid Lind - sigli@itu.dk

Stella Petrova Boneva - stpb@itu.dk

BSPRDAS1KU

May 2025

Abstract

Melanoma is one of the most aggressive forms of skin cancer, where early detection is crucial for improving patient outcomes. This project investigates how extending the traditional ABC features (Asymmetry, Border Irregularity, and Color Variation) with Diameter and Evolution (ABCDE) impacts the performance of machine learning models in classifying melanoma. Analyzing a dermoscopic image dataset containing 2,298 image samples, where class imbalance was addressed using the SMOTE method. Three classifiers (K-Nearest Neighbors, Decision Tree, and Random Forest) were evaluated based on their recall scores. The Decision Tree consistently achieved the highest recall in both the baseline (ABC) and extended (ABCDE) feature sets, making it the preferred model for detecting melanoma. The project results show that adding Diameter and Evolution improved recall performance on the validation set, suggesting that these features capture additional lesion characteristics relevant to melanoma detection. However, the performance on the test set differed from the validation set, underlining the need for further validation on larger, more diverse datasets. This project shows the importance of choosing the right features and testing the model carefully to make skin cancer detection more accurate.

1 Introduction

Melanoma is more dangerous than any other form of skin cancer due to its high potential to spread and affect other parts of the body. Even though the cases of melanoma have been rising for the past decade, the five-year survival rate has also improved, most likely due to improvements in treatment and early detection. The detection of melanoma in the early stage could be crucial for curing it, but the standard methods of diagnosing the deadly cancer (dermatoscopy, biopsy) can be expensive, stressful, and time-consuming.

Therefore, deep learning and machine learning algorithms are being used to analyze lesion images. Different classifiers and computational techniques help to improve the accuracy of a model, which helps determine whether the skin lesion is cancerous. Even though a model that is fully reliable, still hasn't been designed, these approaches show significant progress both in the fields of artificial intelligence and dermatology.

In dermatological practice, the ABCDE rule is widely used and consists of the following clinical features: Asymmetry, Border irregularity, Color variation, Diameter, and Evolving. The ABCDE rule is preferred over the ABC rule, as the diameter and the change of the lesion enhance the early detection of melanoma. This report aims to answer the question: "How does extending ABC features with Diameter and Evolving affect melanoma classification performance?"

2 Dataset analysis

2.1 Dataset

The dataset is collected along with the Dermatological and Surgical Assistance Program at the Federal University of Espírito Santo in Brazil. It consists of 2,298 samples of six different types of skin lesions - three skin cancer types(MEL, BCC, SCC) and three skin diseases(ACK, NEV, SEK). Each sample of the metadata consists of a dermoscopic image and up to 26 clinical features - diameter, gender, skin cancer history, age, etc.

2.2 Data exploration

Melanoma vs. Non-melanoma was chosen because of the high level of danger this skin lesion cancer presents. During the data exploration, there were several key takeaways, used for further analysis. The melanoma cases in the dataset are significantly less than

the non-melanoma ones. Therefore, a binary classification (MEL vs. Non-MEL) is very imbalanced. A resampling technique, SMOTE, is applied to the training set to create new, artificial data points that resemble the minority class, in this case, melanoma.

2.3 Data cleaning

Two images(PAT_488_931_321 and PAT_1725_3222_943), are first noticed because of their negative compactness score in feature extraction, which leads to visual inspection. The findings are that the segmentation masks are inaccurate when it comes to their borders. Especially since they include large non-lesion areas, resulting in a small perimeter and big area which leads to a negative compactness score. To avoid manual data cleaning and make sure the code can run on an external dataset, we return a missing value for the score instead. The absent values are filled with the mean from the feature scores column to avoid losing data. Furthermore, rows in the dataset where the image did not have a corresponding mask are removed, in order to preserve data integrity. After applying the hair removal method, all images with PSNR(Peak Signal-to-noise Ratio) less than 15 and SSIM(Structural Similarity Index Measurement) less than 0.5 are removed. The boundaries were chosen to balance between having a high image quality and a large dataset size. By automated control, the dataset is refined to provide the model with the highest level of consistency.

2.4 Image preprocessing

All images are converted to grayscale, which simplifies edge detection and makes it easier to identify and remove hair using Laplacian edge detection. The method finds hair by detecting regions of rapid intensity changes. Based on the mean color of those edges in the original image, a hair mask is created using either a black-hat operation (for dark hair) or a top-hat operation (for light hair). Using inpaint the hair masks are filled in. This slightly blurs the image but it ensures that unwanted hair and other edges are removed, which could otherwise interfere with feature extraction.

Several additional preprocessing techniques were considered, including bilateral filtering, histogram equalization, and resizing the images to standard 224x224 pixels. However, these were not included in the final implementation. Bilateral filtering made the images too blurry, resulting in a loss of important color details, especially for Feature C. Histogram equalization improved contrast but caused color loss when converting back to RGB. Resizing would have made the model faster but led to a reduction in detail after shrinking the images. These extra techniques were excluded from the final preprocessing model

because of the negative side effects.

3 Feature extraction

3.1 Features

Asymmetry
To compute the asymmetry score the center of the lesion mass is found and cropped. It is then flipped and rotated at a ten-degree angle, chosen because of the faster execution and enough samples compared to five degrees. For each rotation, the lesion is split vertically and horizontally. The horizontal asymmetry (top vs bottom) and the vertical asymmetry(left vs right) are computed by flipping one half and comparing them to the other. The result is normalized, returning a mean asymmetry score between 0(symmetrical) and 1(not symmetrical).

Border irregularity
To estimate the lesion border irregularity a compactness method is used. To implement it, the segmentation masks were used for calculating the area(the white pixels in the mask), and the perimeter(the boundary pixels), applying the formula, and normalizing it. We get an output between 0(smooth) and 1(irregular).

Color
To calculate the color variation score for a lesion the standard deviations are computed for each color channel (R, G, and B) in the lesion. The standard deviations are summed and normalized. A higher score indicates bigger color variation, which increases the probability of melanoma. During data exploration, melanoma showed the highest color variation. The median heterogeneity score for MEL is around 1000 per image, while for others is 400 - 500.

Diameter
This feature is computed as the average of the two given diameter measurements (diameter 1 and diameter 2). This provides a single, consistent estimate of lesion size.

Evolving
This binary feature indicates whether a lesion has changed over time. If a lesion either grew or changed according to the dataset, the feature is set to `True`, suggesting potential

malignancy-related progression.

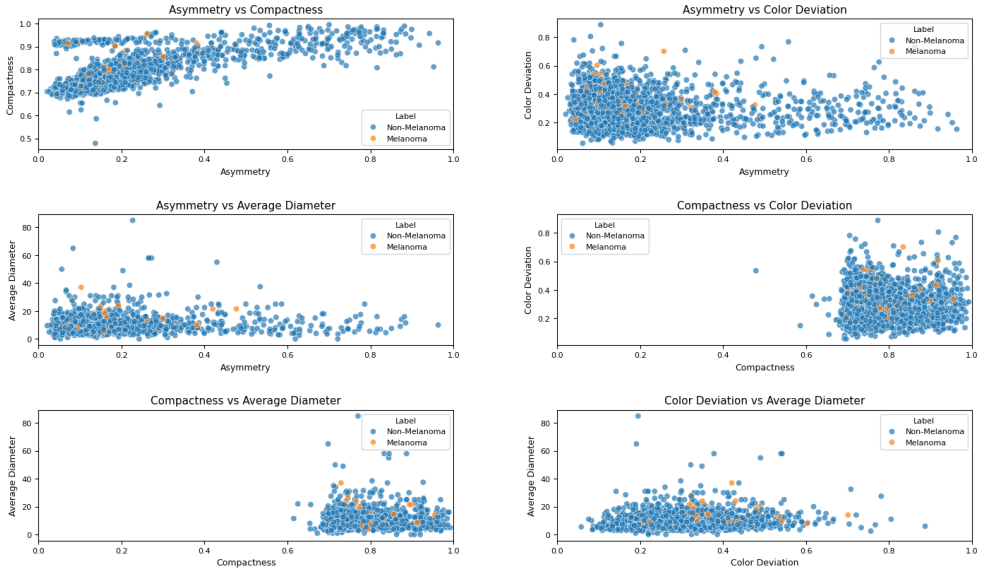


Figure 1: Scatter plots showing relationships between features: asymmetry, compactness (border irregularity), color deviation, and average diameter. Each point represents a lesion, colored by class, non-melanoma(blue) or melanoma(orange).

3.2 Features Post-Processing

Missing values were found for the features Color (3% NaNs), Diameter (32% NaNs), and 2 images for Border irregularity. For each feature column, the missing values were filled with the mean of the column to not drop a large portion of the dataset.

4 Classification

4.1 Splitting the data

The data is split into three sets: training, validation, and test. 70% is used to train the model and the remaining 30% is evenly divided between the validation and test sets. 70/15/15 is chosen not only because it is a default in machine learning tasks. It is preferred over 60/20/20 due to its larger training dataset, which improves the model learning and it is preferred over 80/10/10 because 10% validation and test sets might be too small to provide reliable results.

4.2 SMOTE

The dataset consists of 2246($\sim$ 98%) Non-MEL cases and 52($\sim$ 2%) MEL cases, due to the large imbalance in our binary classification SMOTE is applied (Synthetic Minority Oversampling Technique) to the training set. The reason this method was chosen is that it handles class imbalances in classification by creating synthetic samples of the minority class. SMOTE is only applied on the training set and not validation or test to avoid leaking of synthetic data which could result in an overestimated performance, due to distributional overlap.

4.3 Selecting Classifiers

To classify whether a lesion is melanoma or not, three different machine learning models are used: K-nearest neighbors (KNN), Decision Tree (DT), and Random Forest (RT). These are selected since they can handle non-linear relationships, which are expected in features like asymmetry and color variation (Figure 1).

4.4 Tuning model

The parameters for the chosen models (KNN, DT, RF) are tuned based on the Recall score. This metric was chosen because true melanoma cases(false negatives) can be extremely dangerous by delaying treatment and leading to more complications. While the F1 score balances between recall and precision, in a medical context missing a melanoma case(false negative) is much more critical than a false alarm. Recall was therefore chosen.

5-Fold Cross-Validation is applied when training each model, five splits because it is a balance between computational efficiency (more folds) and low variance in validation estimates. Each model is trained and validated five times since it is partitioned into that amount of subsets. This means that each time a different fold is used as the validation set the risks of overfitting to a particular split are reduced, while resulting in a more robust estimate of model performance.

4.5 Choosing final classifier

Before analyzing the performance of each classifier, it is important to clarify the difference between the baseline and the extended research. In the baseline research, only the ABC features were used - these correspond to Asymmetry, Border Irregularity, and Color. In

the extended research, two additional features, Diameter and Evolving, were added to the feature set (ABCDE). This extension was intended to capture more nuanced characteristics of the lesions, potentially improving the classifier's ability to distinguish melanoma from non-melanoma cases.

To determine the most appropriate model for melanoma detection, we compare the performance of the trained classifiers using recall, although we have several metrics given. In the context of melanoma detection, recall represents the model's ability to correctly identify positive cases (melanomas), thereby reducing the risk of false negatives.

4.6 Evaluating result to choose classifier

In the baseline models, the Decision Tree achieved the highest recall (0.500) in the validation set, which is why the model was chosen.

	Precision	**Recall**	**F1-score**	**Support**
Melanoma	0.05	0.83	0.10	6
Non-Melanoma	1.00	0.70	0.82	309
Macro avg	0.52	0.77	0.46	315
Weighted avg	0.98	0.70	0.81	315

Figure 2: Baseline Test set Performance: Classification report showing precision, recall, F1-score, and support for Non-Melanoma and Melanoma classes.

In the extended models, the Decision Tree again achieved the highest recall (0.600) on the validation set. Notably, all the extended models performed better on F1-Score and AUC than their baseline counterparts, reflecting the benefits of incorporating the additional Diameter and Evolving features. However, in the test set baseline performs better in Recall for melanoma cases whereas the extended performs better on average. This could be due to a very small amount of melanoma cases.

	Precision	**Recall**	**F1-score**	**Support**
Melanoma	0.04	0.50	0.08	6
Non-Melanoma	0.99	0.79	0.88	309
Macro avg	0.52	0.65	0.48	315
Weighted avg	0.97	0.79	0.86	315

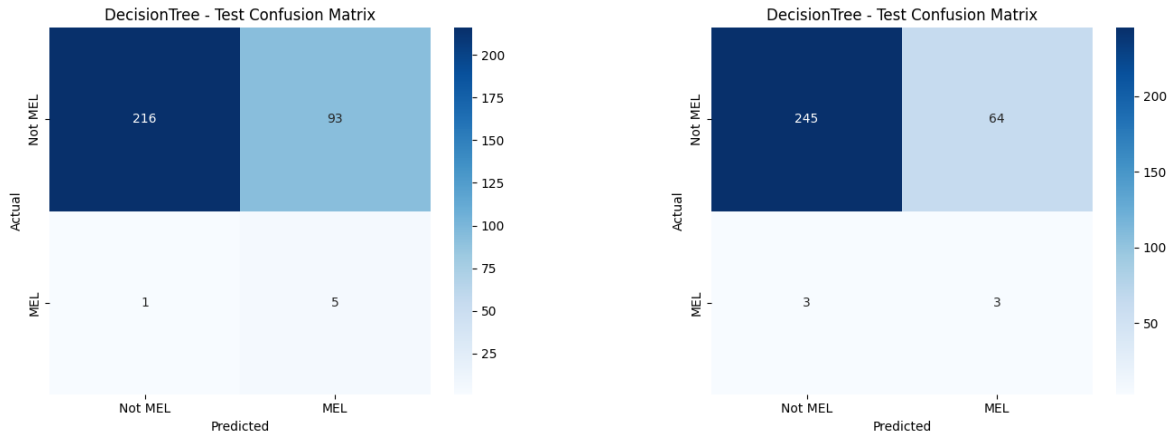
Figure 3: Extended Test set Performance: Classification report showing precision, recall, F1-score, and support for Non-Melanoma and Melanoma classes.

Given the critical importance of recall in medical diagnostics, the Decision Tree was

selected as the final classifier for both the baseline and extended approaches, based on its highest recall score in both setups.

On the test set, the baseline Decision Tree achieved a higher recall (0.833) compared to the extended Decision Tree (0.500). This difference may be due to several factors. First, the test set contains a very small number of melanoma cases, making the recall metric more sensitive to even a single correct or incorrect prediction. For example, correctly identifying one additional melanoma case can significantly increase the recall percentage. Secondly, adding more features (ABCDE) leads to increased model complexity and possibly capturing noise.

However, the extended model consistently showed stronger performance on the validation set, which suggests that the additional features (ABCDE) could contribute valuable information. Therefore, the higher test recall in the baseline model may simply reflect variability due to limited sample size rather than a definitive indication that it is a generally better model.



(a) Baseline model

(b) Extended model

Figure 4: Confusion matrices for the baseline and extended models using the Decision Tree classifier. The matrices show the classification results for melanoma (MEL) and non-melanoma (Not MEL) cases on the test set.

5 Limitations

Limitations can be observed in our model. The dataset shows a difference in hair color, hair amount, lighting, and image quality. However, it can be a problem for the model accuracy if it is trained on non-diverse data, hereby a bias is introduced into the model performance. The model has only been trained on a finite skin tone. Therefore, using external images with a different variety of skin tones may have improved bias in the model.

Missing values were observed during data exploration mainly in diameter (32%). This high proportion of missing values could affect the consistency and reliability of classification. After feature extraction other missing values were observed, caused by filtering based on image quality after hair removal. Images of minimal hair on the lesion would have prevented loss of quality in the image.

The accuracy of lesion segmentation is extremely important for feature extraction, in particular Asymmetry and Border irregularity. By visual inspection variability in mask quality was found, which introduces noise in the extracted features, potentially affecting the performance. Images without corresponding masks were removed which limits our data samples. An imbalance was present in the dataset, therefore SMOTE was applied. This method has limitations, mainly that synthetic samples may not fully represent real-world variability, potentially leading to overfitting and reduced model generality.

6 Conclusion

6.1 Final results

Based on our analysis, the baseline model had a higher recall on the test set, this is likely due to few melanoma cases and random variation, rather than true robustness. Incorporating Diameter and Evolving features (ABCDE) in addition to the ABC features resulted in improved model performance on the validation set, particularly in the recall, which is critical for melanoma detection. Overall, our results suggest that extending the ABC features with Diameter and Evolving can enhance melanoma classification performance by capturing more nuanced lesion characteristics.

6.2 Future Work

Our model used standard k-fold cross-validation, which splits data randomly into folds for training and validation. Since our dataset is imbalanced, this could lead to some folds with few or no melanoma cases, making validation results inconsistent. Using stratified k-fold cross-validation instead would ensure that each fold retains the same class balance as the whole dataset, leading to more reliable validation scores and potentially improving model performance. In general, we could have added more features i.e. blue veil, globules, irregular pigmentation, streaks, and vascular.

References

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". "Journal of Artificial Intelligence Research", no. 16 (2002): 321–357.

American Cancer Society. "Cancer Facts and Figures 2025". (2025): 22–24.

Arslan Javaid, Muhammad Sadiq, Faraz Akram. "Skin Cancer Classification Using Image Processing and Machine Learning". (2021)

Detection and Classification of Melanoma Skin Cancer Using Image Processing Technique